Diagramas de dispersion en el análisis de datos

In [ ]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

#Configurando Latex
# Configuración de Matplotlib para usar LaTeX
plt.rcParams.update({
    "text.usetex": True,
    "font.family": "serif",
    "font.serif": ["Computer Modern Roman"],
    "text.latex.preamble": r"\usepackage{amsmath}"
})

iris = sns.load_dataset('iris')
#Dataset de Flores
iris.head(3)
```

Out[ ]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa

El data set tiene atributos que se miden en la flor Iris y entonces tenemos que dependiendo de la longitud de 2 elementos fundamentales; pétalo y sepalo podemos clasificar a que especie corresponden.

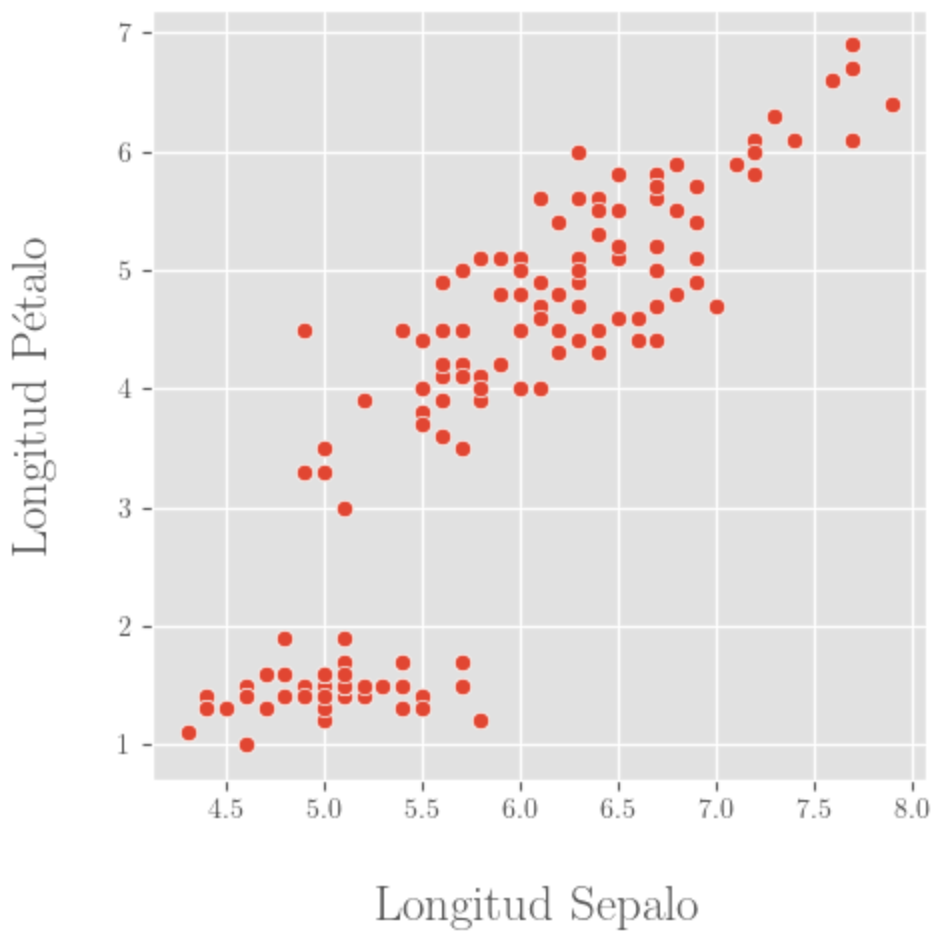
Vamos a trabajar con la gráfica scatterplot.

In [ ]:
```
iris.columns
```

Out[ ]:
```
Index(['sepal_length', 'sepal_width', 'petal_length', 'petal_width',
       'species'],
      dtype='object')
```

In [ ]:
```
#Definiendo estilos
plt.style.use('ggplot')
fig, ax = plt.subplots(figsize=(5,5))

sns.scatterplot(data=iris,x='sepal_length',y='petal_length')
plt.title('Relación\nSepalo vs Pétalo\n',fontsize=23)
plt.xlabel('\nLongitud Sepalo',fontsize=18)
plt.ylabel('Longitud Pétalo\n',fontsize=18)
```

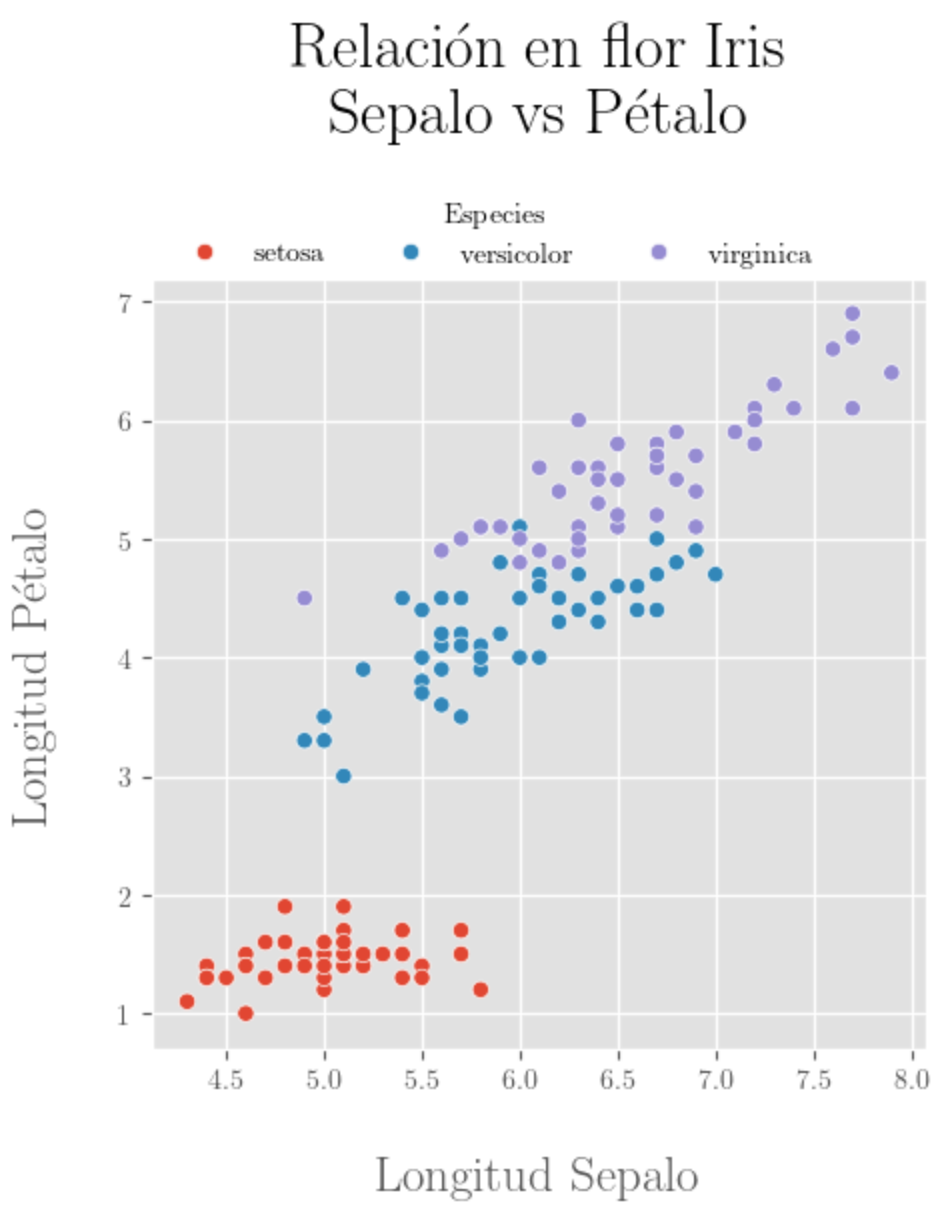
Out[ ]:
```
Text(0, 0.5, 'Longitud Pétalo\n')
```



Esta gráfica dibuja un punto por flor y la intersección entre ambas medidas **(longitud del sepalo y longitud del pétalo)** es donde se ubica el punto. Este tipo de visualización nos muestra la relación de los datos. Aquí están combinadas todas las especies.

Hagamos una segmentación de especies.

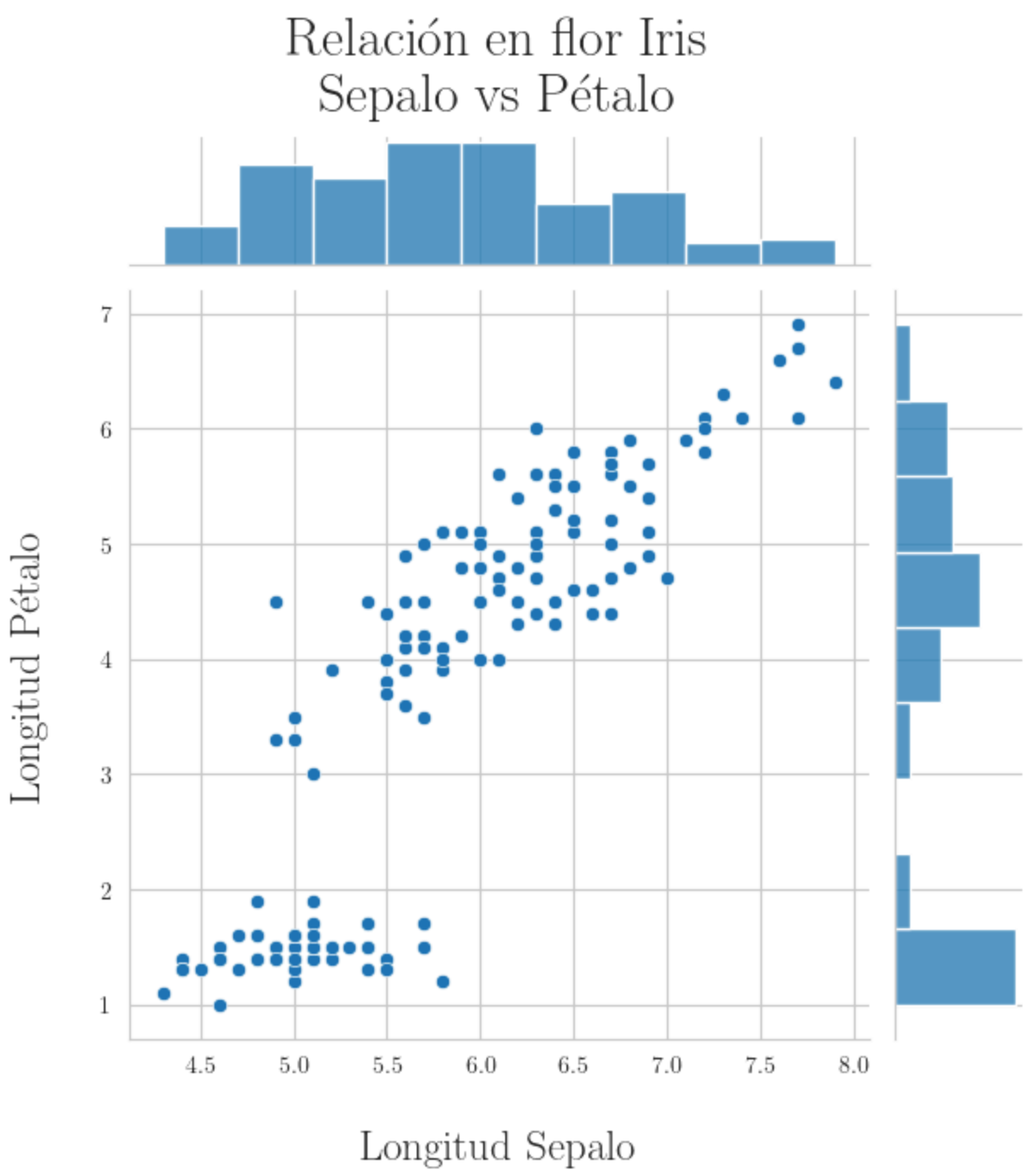
In [ ]:
```
#Definiendo estilos
plt.style.use('ggplot')
fig, ax = plt.subplots(figsize=(5,5))

sns.scatterplot(data=iris,x='sepal_length',y='petal_length',hue='species')
plt.title('Relación en flor Iris\nSepalo vs Pétalo\n\n',fontsize=23)
plt.xlabel('\nLongitud Sepalo',fontsize=18)
plt.ylabel('Longitud Pétalo\n',fontsize=18)
sns.move_legend(ax, "upper left", bbox_to_anchor=(0, 1.13),ncol=3,title='Especies',frameon=False)
```



Aquí en este gráfico se percibe una relación mas clara sobre los 3 tipos de especies y el tamaño de sus parámetros, por ejemplo podemos observar que las tipo **setosa** tienen la menor *longitud en sepalo* y de igual forma la menor *longitud de pétalo*.

A continuación usaremos Joinplot.

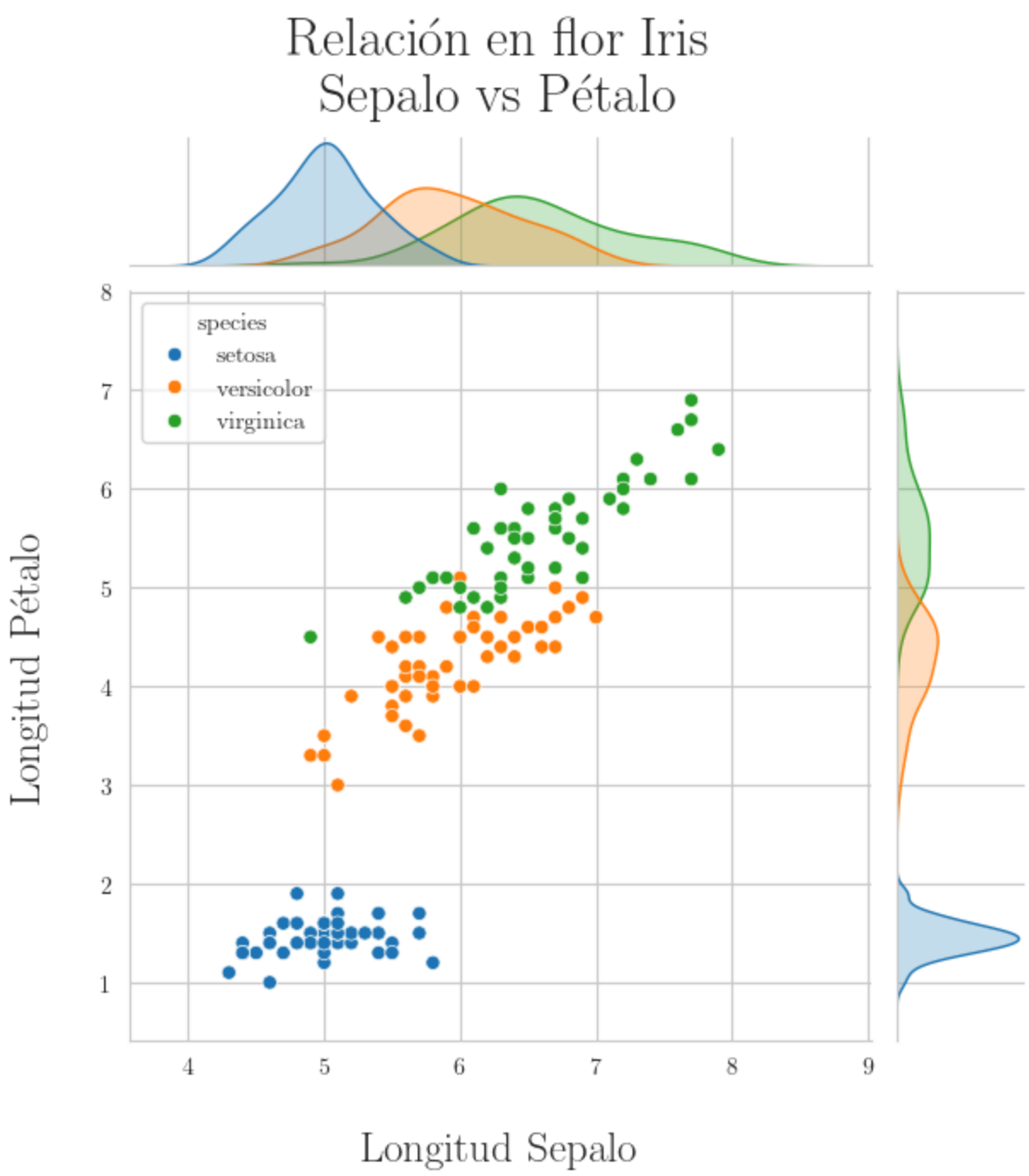
In [ ]:
```
#Definiendo estilos
plt.style.use('default')
plt.rcParams.update({
    "text.usetex": True,
    "font.family": "serif",
    "font.serif": ["Computer Modern Roman"],
    "text.latex.preamble": r"\usepackage{amsmath}"
})

with sns.axes_style("whitegrid"):
    sns.jointplot(data=iris,x='sepal_length',y='petal_length')
    plt.title('Relación en flor Iris\nSepalo vs Pétalo\n\n\n',fontsize=23)
    plt.xlabel('\nLongitud Sepalo',fontsize=18)
    plt.ylabel('Longitud Pétalo\n',fontsize=18)
```



Como se puede observar este tipo de visualización combina 2 gráficas, en este caso arrojó un scatterplot y 2 histogramas para cada eje.

Ahora agregaremos una segmentación de especies.

In [ ]:
```
#Definiendo estilos
plt.style.use('default')
plt.rcParams.update({
    "text.usetex": True,
    "font.family": "serif",
    "font.serif": ["Computer Modern Roman"],
    "text.latex.preamble": r"\usepackage{amsmath}"
})

with sns.axes_style("whitegrid"):
    sns.jointplot(data=iris,x='sepal_length',y='petal_length',hue='species')
    plt.title('Relación en flor Iris\nSepalo vs Pétalo\n\n\n',fontsize=23)
    plt.xlabel('\nLongitud Sepalo',fontsize=18)
    plt.ylabel('Longitud Pétalo\n',fontsize=18)
```



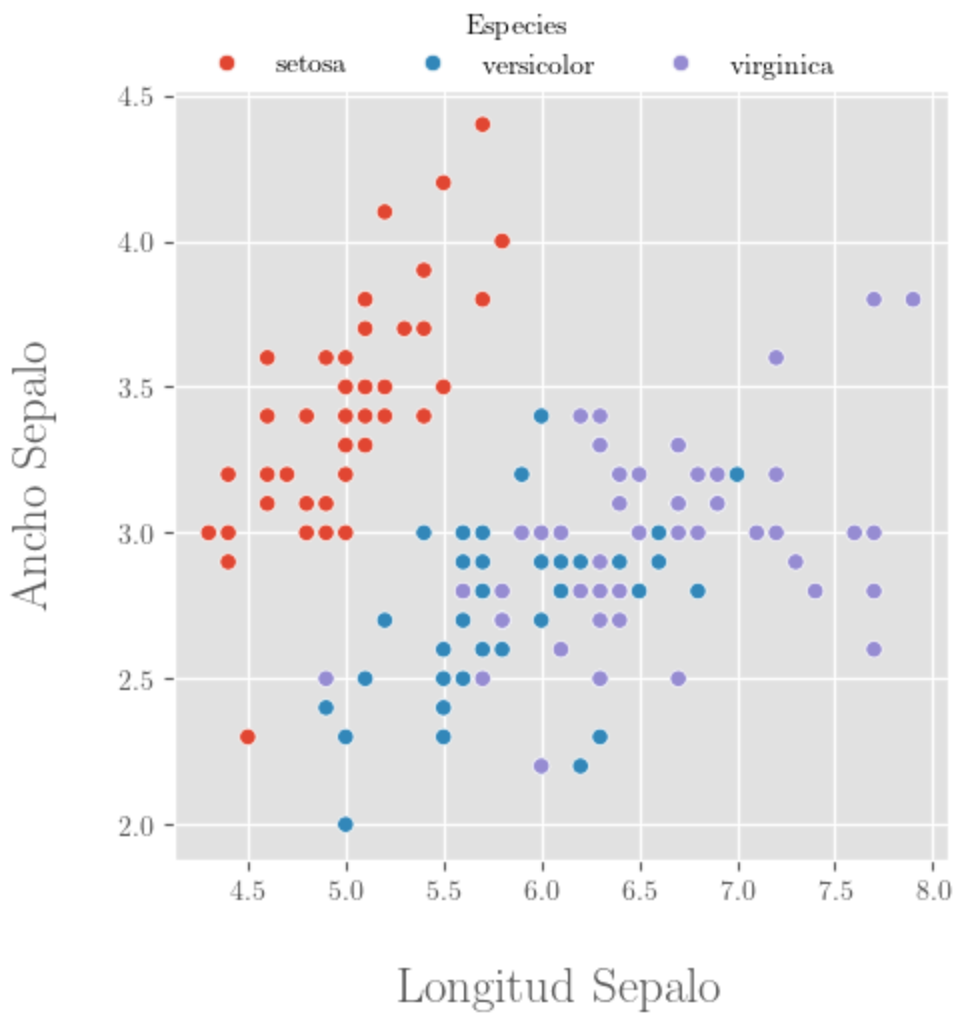
Como podemos observar, aquí se cambio el tipo de gráfica en los ejes y esto es debido a la segmentación. También podemos apreciar en los diferentes ejes que las distribuciones de los diversos datos, en algunos casos pueden parecer uniformes y en otros no parecen asi.

También podemos comparar las demás variables.

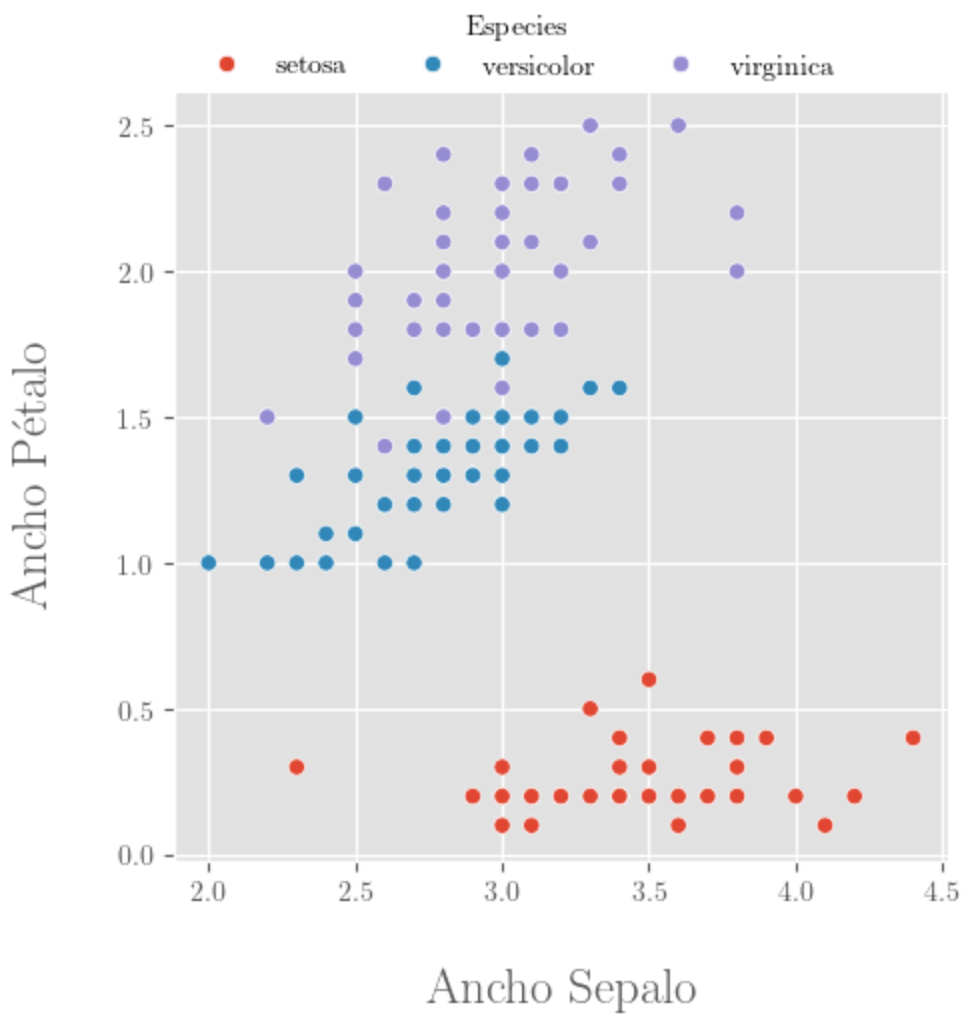
In [ ]:
```
#Veremos que columnas disponibles tenemos
iris.columns
```

Out[ ]:
```
Index(['sepal_length', 'sepal_width', 'petal_length', 'petal_width',
       'species'],
      dtype='object')
```

In [ ]:
```
#Definiendo estilos
plt.style.use('ggplot')
fig, ax = plt.subplots(figsize=(5,5))

sns.scatterplot(data=iris,x='sepal_length',y='sepal_width',hue='species')
plt.title('Relación en flor Iris\nSepalo\n\n',fontsize=23)
plt.xlabel('\nLongitud Sepalo',fontsize=18)
plt.ylabel('Ancho Sepalo\n',fontsize=18)
sns.move_legend(ax, "upper left", bbox_to_anchor=(0, 1.13),ncol=3,title='Especies',frameon=False)
```

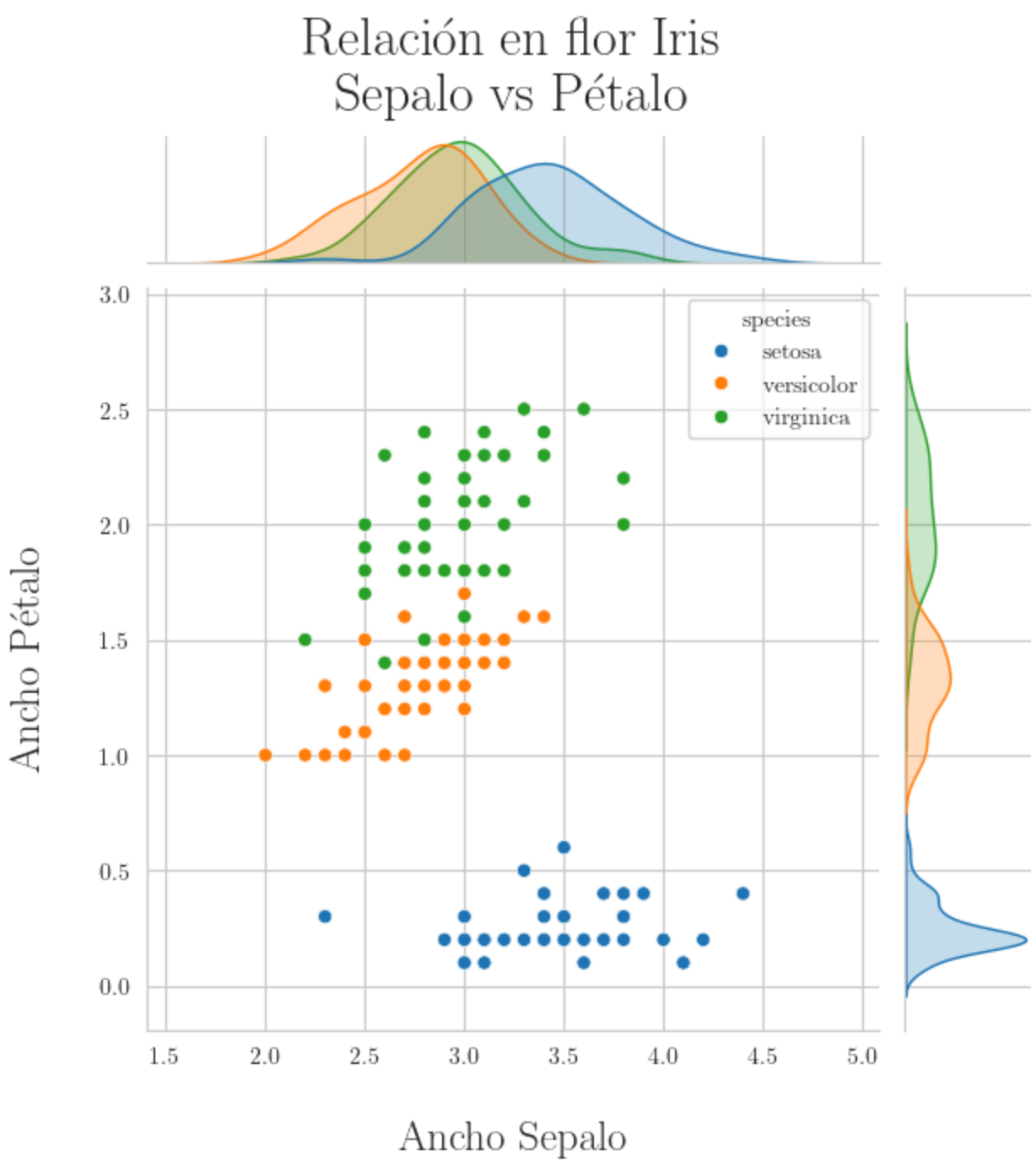


In [ ]:
```
#Definiendo estilos
plt.style.use('ggplot')
fig, ax = plt.subplots(figsize=(5,5))

sns.scatterplot(data=iris,x='sepal_width',y='petal_width',hue='species')
plt.title('Relación en flor Iris\nSepalo vs Pétalo\n\n',fontsize=23)
plt.xlabel('\nAncho Sepalo',fontsize=18)
plt.ylabel('Ancho Pétalo\n',fontsize=18)
sns.move_legend(ax, "upper left", bbox_to_anchor=(0, 1.13),ncol=3,title='Especies',frameon=False)
```



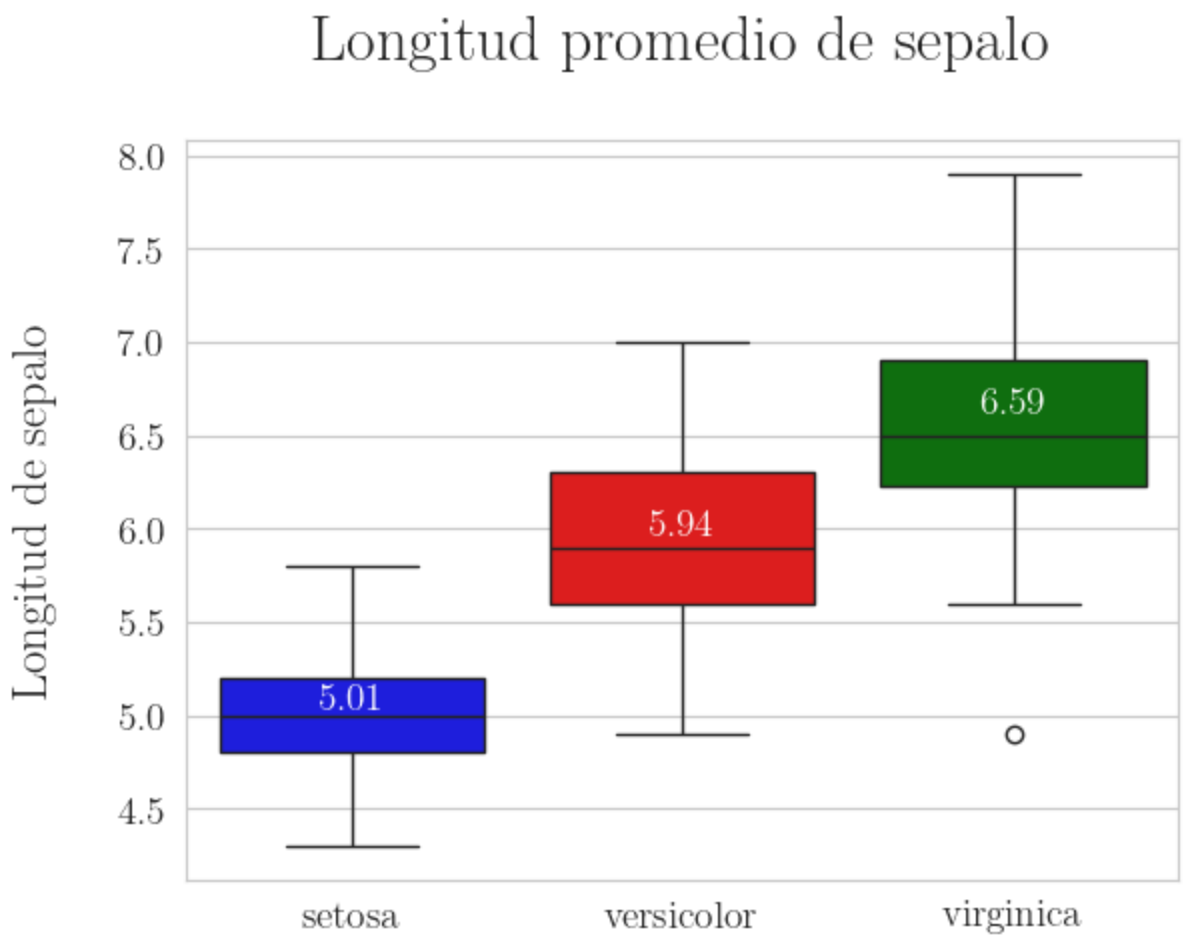
In [ ]:
```
#Definiendo estilos
plt.style.use('default')
plt.rcParams.update({
    "text.usetex": True,
    "font.family": "serif",
    "font.serif": ["Computer Modern Roman"],
```

```
    "text.latex.preamble": r"\usepackage{amsmath}"
})

with sns.axes_style("whitegrid"):
    sns.jointplot(data=iris,x='sepal_width',y='petal_width',hue='species')
    plt.title('Relación en flor Iris\nSepalo vs Pétalo\n\n\n',fontsize=23)
    plt.xlabel('\nAncho Sepalo',fontsize=18)
    plt.ylabel('Ancho Pétalo\n',fontsize=18)
```



In [ ]:
```
#Box plot
with sns.axes_style("whitegrid"):
    sns.boxplot(x='species',y='sepal_length',data=iris,palette=['blue','red','green'])
    plt.title('\nLongitud promedio de sepalo\n',fontsize=23)
    plt.xlabel('\nEspecies',fontsize=18)
    plt.ylabel('Longitud de sepalo\n',fontsize=18)

    for index, row in Species_averages_sl.iterrows():
        plt.text(index, row['sepal_length'] + .03, f'{row["sepal_length"]:.2f}', color='white', ha="center")
```

```
/tmp/ipykernel_4506/1246751108.py:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `h
ue` and set `legend=False` for the same effect.

  sns.boxplot(x='species',y='sepal_length',data=iris,palette=['blue','red','green'])
```



Recordemos el uso de cuartiles Q2=mediana, entonces podemos decir que, **la mediana de la longitud del sepalo en**:

- Setosa es 5.0
- Versicolor es 5.9
- Virginica es 6.4

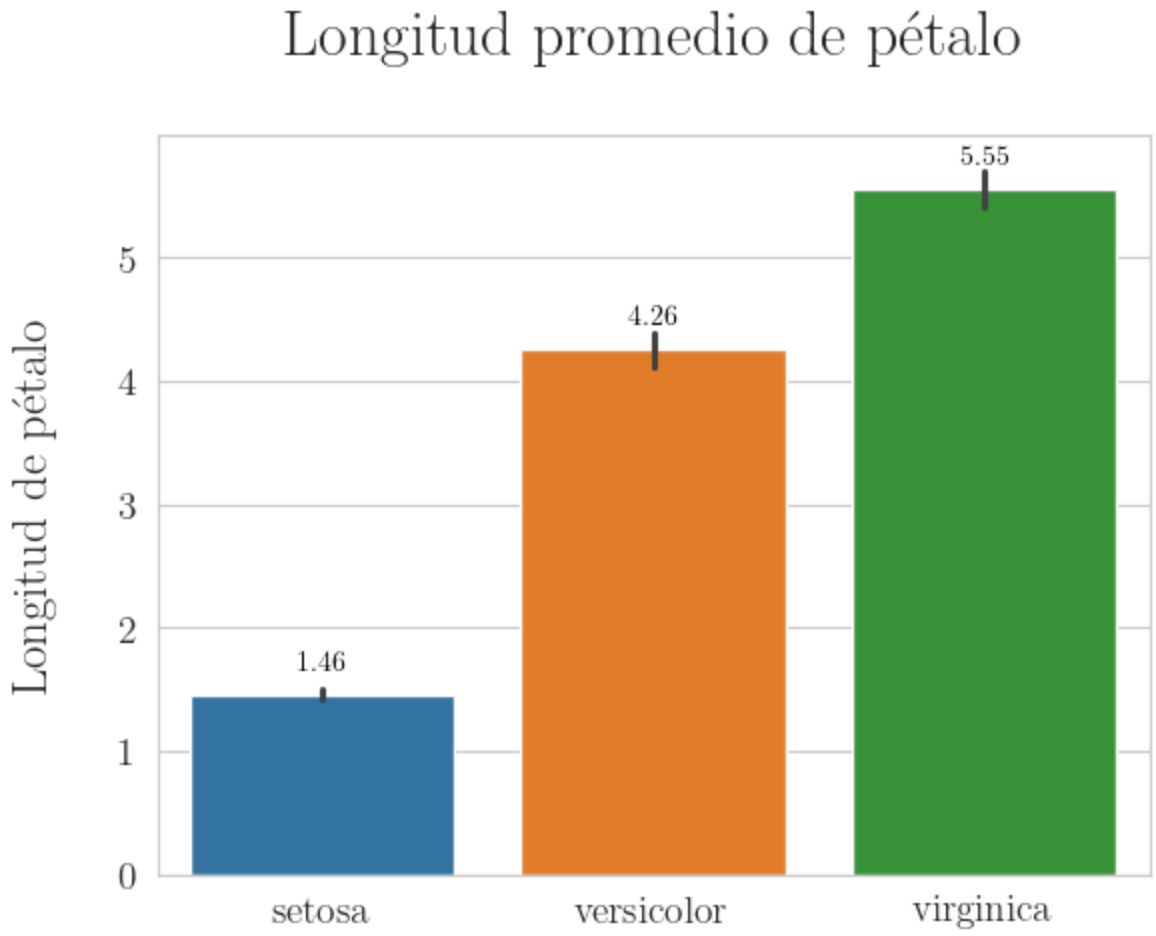
In [ ]:
```
Species_averages_sl = iris.groupby(by='species')['sepal_length'].mean().reset_index()
Species_averages_sw = iris.groupby(by='species')['sepal_width'].mean().reset_index()
Species_averages_pl = iris.groupby(by='species')['petal_length'].mean().reset_index()
Species_averages_pw = iris.groupby(by='species')['petal_width'].mean().reset_index()
iris.groupby(by='species').mean()
```

Out[ ]:

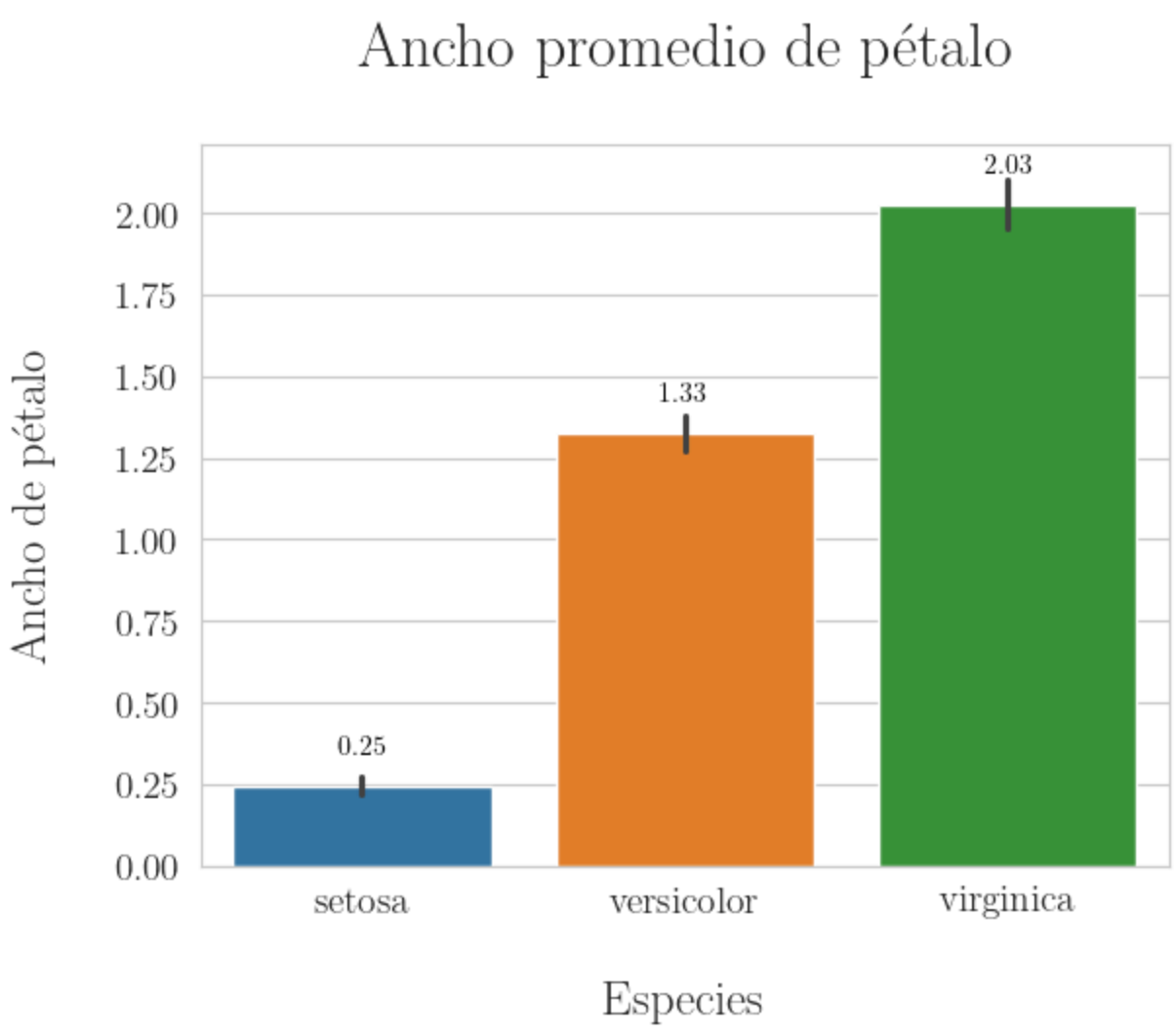
species	sepal_length	sepal_width	petal_length	petal_width
setosa	5.006	3.428	1.462	0.246
versicolor	5.936	2.770	4.260	1.326
virginica	6.588	2.974	5.552	2.026

En la tabla anterior podemos afirmar nuestra hipótesis que teníamos sobre los promedios en **longitud del sepalo de cada especie**

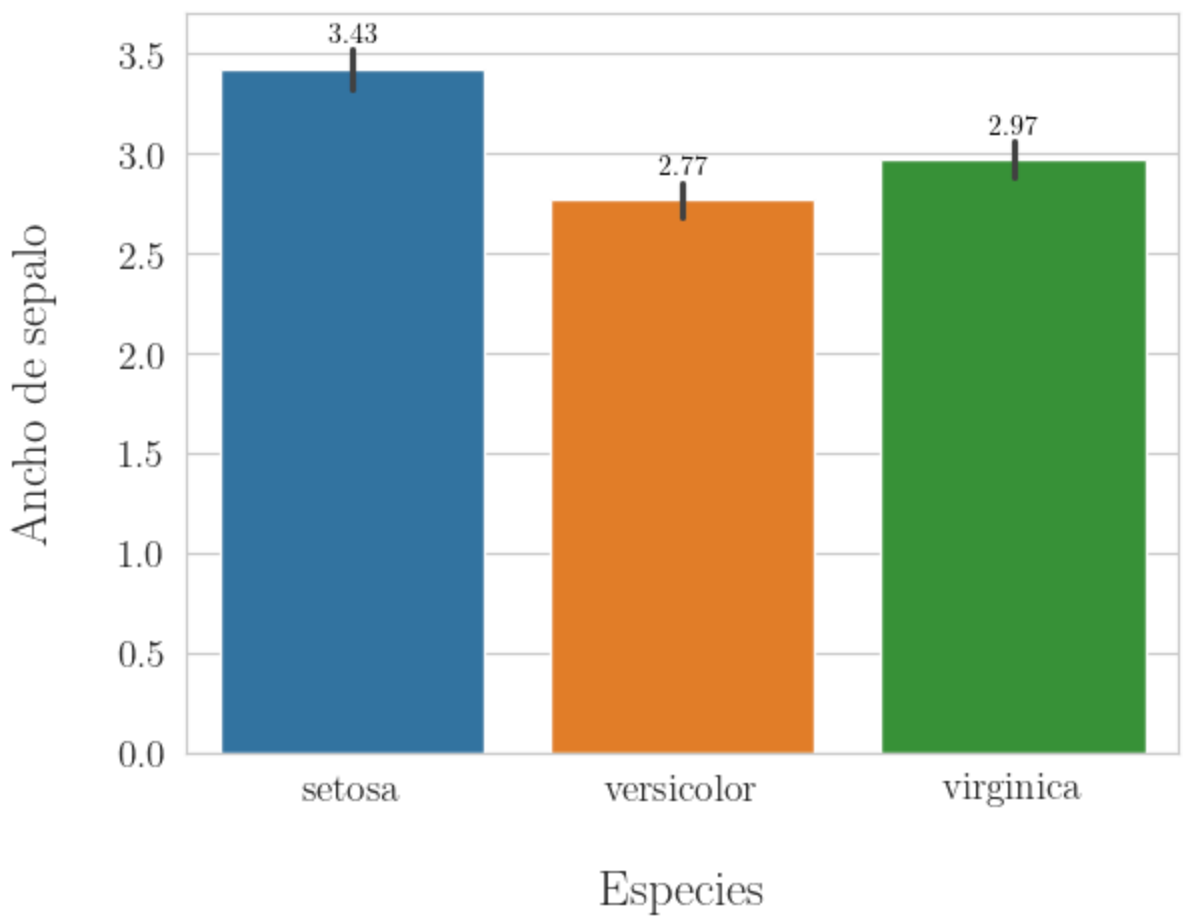
In [ ]:
```
#Longitud promedio de pétalo
with sns.axes_style("whitegrid"):
    sns.barplot(x='species',y='petal_length',data=iris,hue='species')
    plt.title('\nLongitud promedio de pétalo\n',fontsize=23)
    plt.xlabel('\nEspecies',fontsize=18)
    plt.ylabel('Longitud de pétalo\n',fontsize=18)
    for index, row in Species_averages_pl.iterrows():
        plt.text(index, row['petal_length'] + .2, f'{row["petal_length"]:.2f}', color='black', ha="center",fontsize=1
```



In [ ]:
```
#Ancho promedio de pétalo
with sns.axes_style("whitegrid"):
    sns.barplot(x='species',y='petal_width',data=iris,hue='species')
    plt.title('\nAncho promedio de pétalo\n',fontsize=23)
    plt.xlabel('\nEspecies',fontsize=18)
    plt.ylabel('Ancho de pétalo\n',fontsize=18)
    for index, row in Species_averages_pw.iterrows():
        plt.text(index, row['petal_width'] + .1, f'{row["petal_width"]:.2f}', color='black', ha="center",fontsize=10
```



In [ ]:
```
#Ancho promedio de sepalo
with sns.axes_style("whitegrid"):
    sns.barplot(x='species',y='sepal_width',data=iris,hue='species')
    plt.title('\nAncho promedio de sepalo\n',fontsize=23)
    plt.xlabel('\nEspecies',fontsize=18)
    plt.ylabel('Ancho de sepalo\n',fontsize=18)

    for index, row in Species_averages_sw.iterrows():
        plt.text(index, row['sepal_width'] + .13, f'{row["sepal_width"]:.2f}', color='black', ha="center",fontsize=1
```



Extras

https://seaborn.pydata.org/tutorial/properties.html#text